Contents

QBS 103 Final Project

Ayei Chang

August 2025

1 Introduction

The following figures were generated from two datasets from the large-scale multi-omic analysis of COVID-19 severity (Overmyer et al., Cell Systems 2021)[3]. RNA sequencing and high-resolution mass spectrometry were performed on 126 blood samples from COVID-19-positive and COVID-19-negative patients with various disease severities and outcomes to create a reference data set. 219 molecular features with high significance for COVID-19 status and severity were mapped to create the second reference data set [3]. Of the genes in these data sets, the A1BG gene, a member of the immunoglobulin superfamily and a biomarker of various diseases, was chosen to be compared with age, sex, and disease status of the study samples [1].

2 Methods

Gene expression data from the GSE157103 dataset and the clinical metadata from the study were imported into R (Version 2025.05.1+513) using the readr [6] package. The gene expression dataset and the clinical data were merged into one data frame. Data wrangling and visualization were performed with ggplot2 [4] and dplyr [5] packages. A histogram, scatterplot, box plot, and violin plot were generated to visualize the relationship between A1BG gene expression and age, sex, and disease status.

For the heatmap, hierarchical cluster was used in the pheatmap [2] package in R. The expression of ten genes (A1BG, A1CF, A2M, A2ML1, A3GALT2, A4GNT, AAAS, AACS, AADAC, and AADACL2) from all participants in the study were used in the heatmap. The disease status and sex of each participant were also included as annotations to compare the relativeness of the clustering patterns.

A summary statistics table stratified by ICU status was calculated by utilizing the dyplr [5], knitr [7], and kableExtra [8] packages for a LaTeX output. The table includes the prevalence of categorical variables, sex and disease status, as well as the mean and standard deviations of continuous variables, age, A1BG expression, and Charlson score for 120 samples from the study data frame.

3 Results

3.1 Histogram

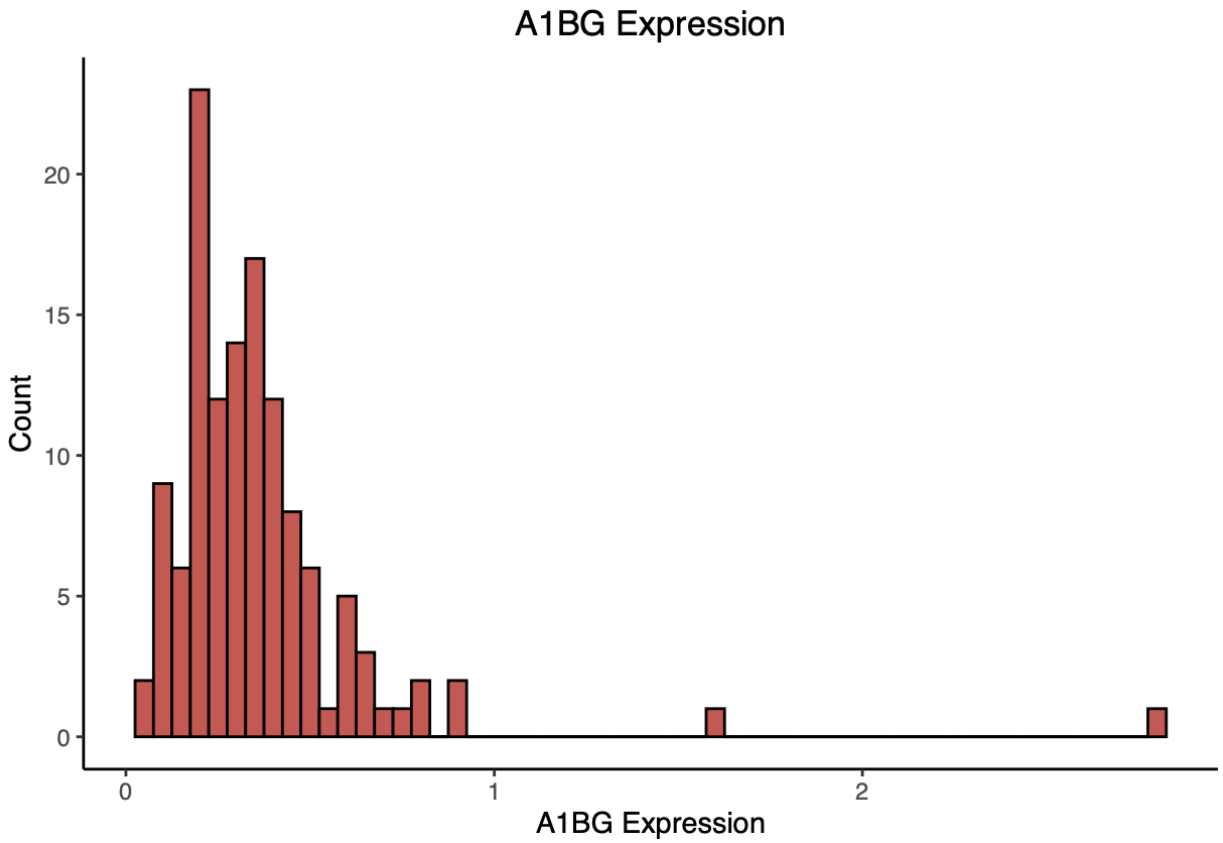

Figure 1: Histogram displaying A1BG gene expression.

This histogram indicates that the A1BG expression of most samples is concentrated between 0.2 and 0.5 (Figure 1). The x-axis is the A1BG expression values, and the y-axis is the number of samples. Almost all samples have expression values between 0 and 1, with the exception of two outliers. The distribution is right-skewed.

3.2 Scatterplot

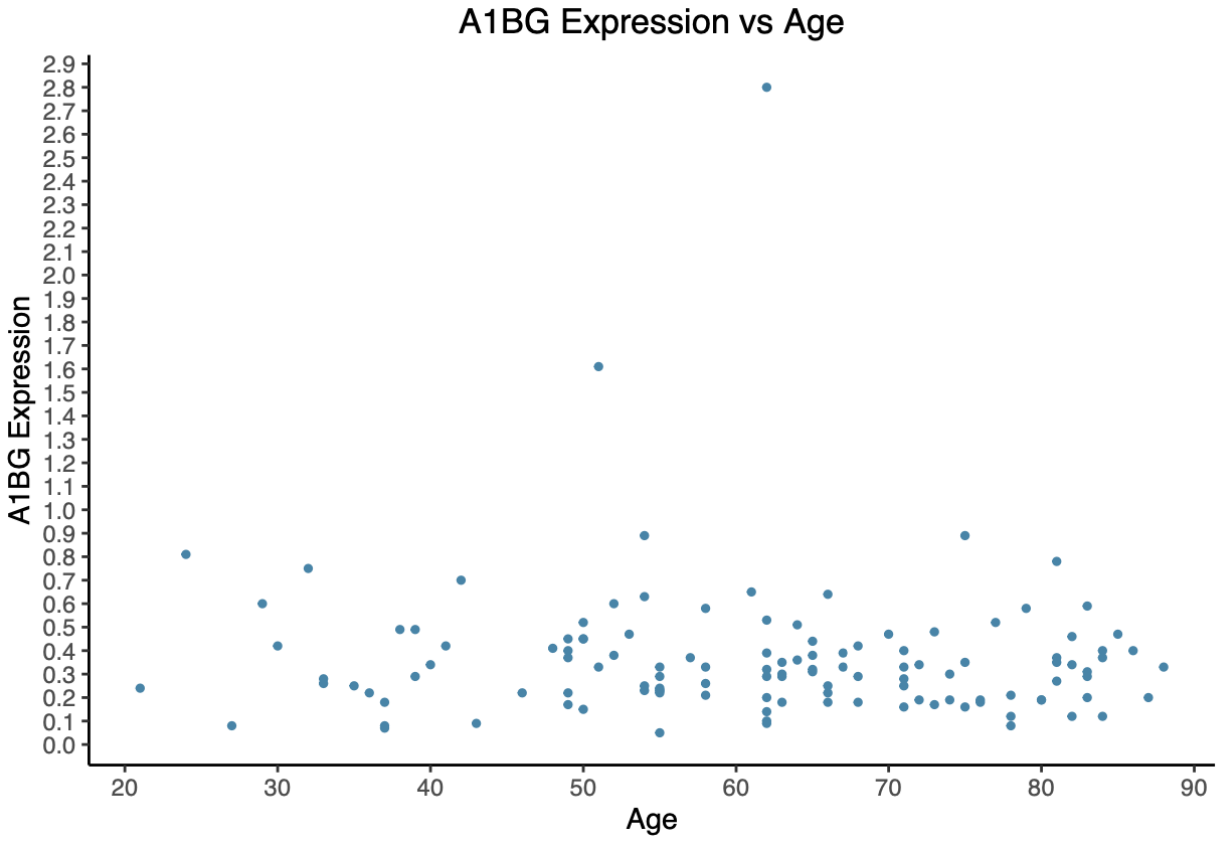

Figure 2: Scatterplot comparing A1BG gene expression by age.

The x-axis displays the age, 20-90 years of age, of the participants, and the y-axis shows the expression values of A1BG. Each point in the scatterplot is a sample plotted by their age and A1BG expression value. There is no strong linear relationship, suggesting that there is no clear association between A1BG expression and age (Figure 2). Across all ages, the gene expression is low, between 0 and 0.6, with two obvious outliers and a couple between 0.7 and 0.9.

3.3 Boxplot

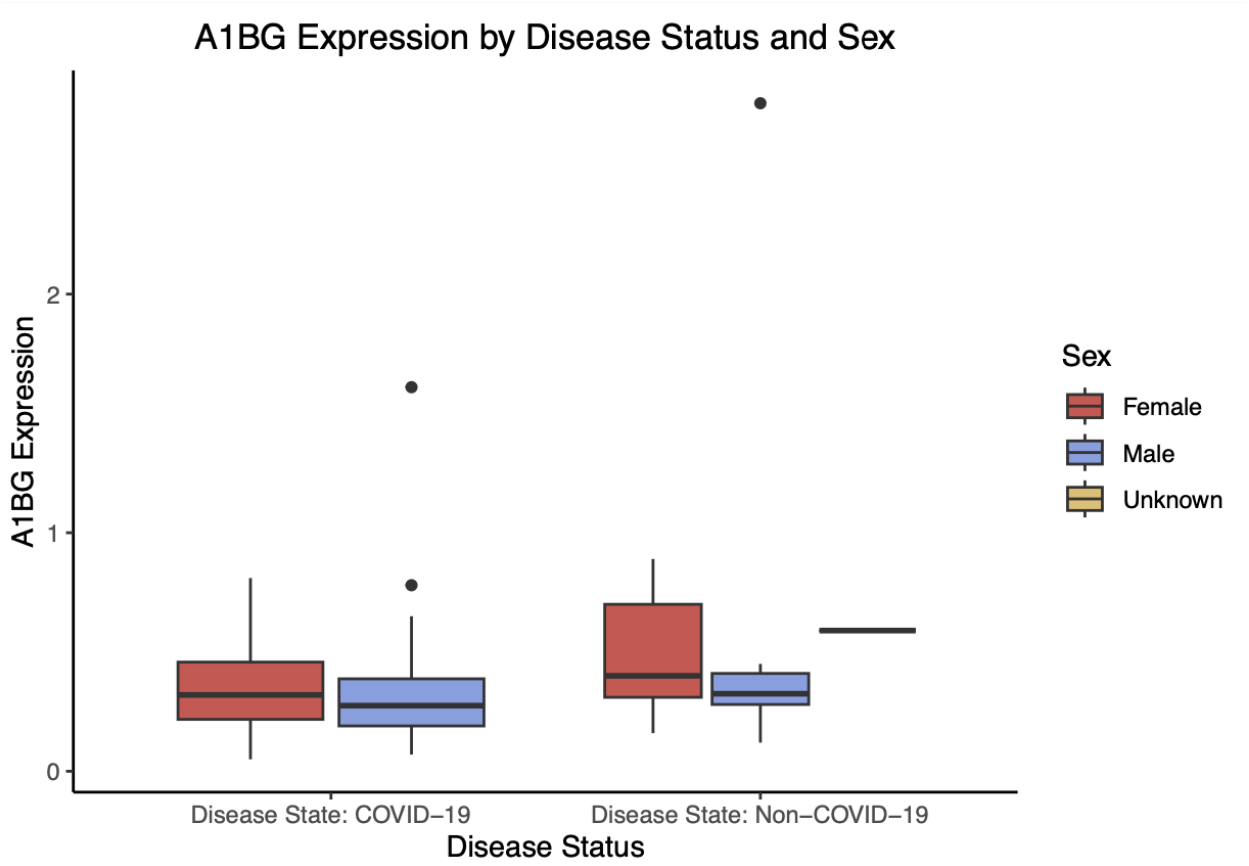

Figure 3: Boxplot comparing A1BG gene expression by disease state (positive or negative for COVID-19) and sex.

The box plot (Figure 3) is divided into two subgroups according to the disease status, as noted on the x axis based on COVID-19 samples and non-COVID-19 samples. Each group is also divided by sex, male and female. The y-axis shows the A1BG gene expression values. In COVID-19 positive samples, the medians of A1BG expression between female and male samples are almost the same, but in COVID-19 negative samples, females have a slightly higher median than males. The spread suggests that A1BG expression is relatively consistent across all groups with 3 outliers in the samples.

3.4 Heatmap

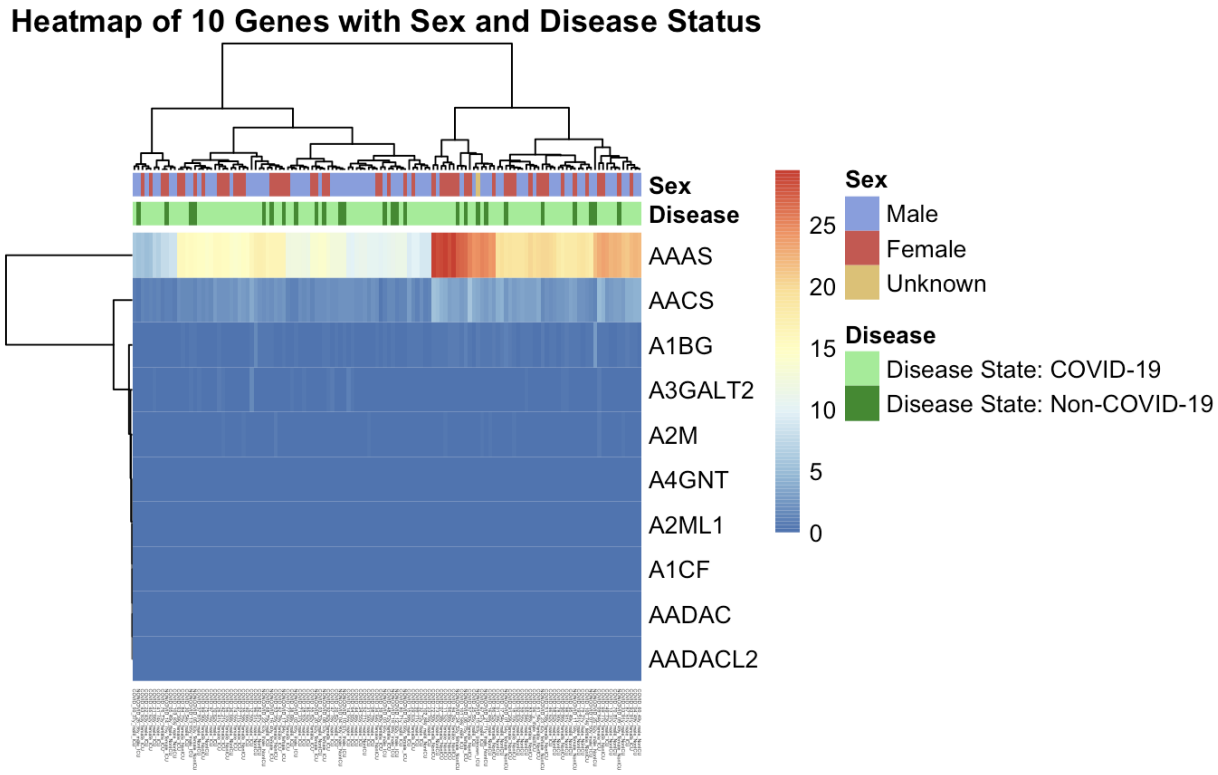

Figure 4: Heatmap displaying the expression of ten different genes across all study participants with sex and disease status tracking bars.

The heatmap (Figure 4) displays ten different gene expressions (A1BG, A1CF, A2M, A2ML1, A3GALT2, A4GNT, AAAS, AACS, AADAC, and AADACL2) for all 120 samples from the study. Sex and disease status are on top as tracking bars. Clustering was used to group samples with similar expressions. The heatmap indicates that there is low expression across the ten genes in the study, with AAAS being the highest. There is no significant relationship between any of the gene expressions with sex or disease status.

3.5 Violin Plot

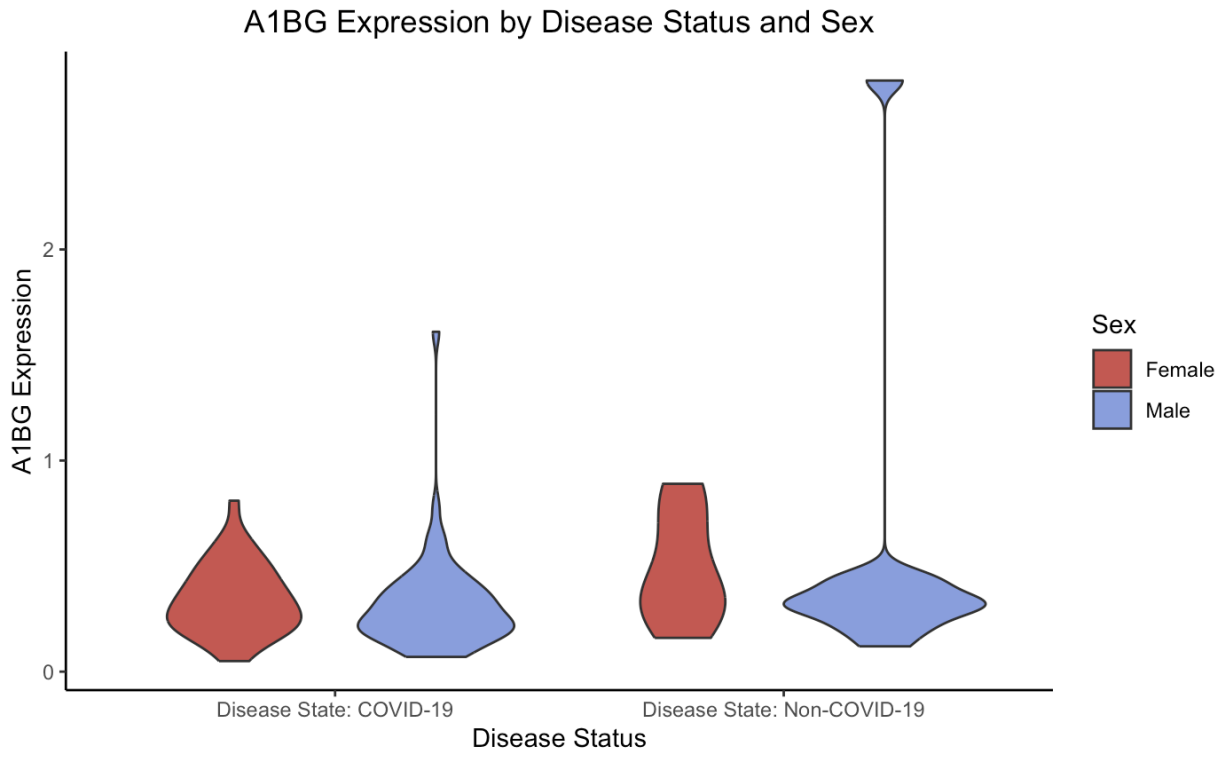

Figure 5: Violin plots displaying A1BG Expression by disease status (positive or negative for COVID-19) and sex.

The violin plot (Figure 5) is divided by disease status (as shown on the x-axis) and further by sex (red for female, blue for male). The y-axis shows the A1BG expression. Each violin plot indicates the density of expression for each subgroup. The distribution of A1BG expression is relatively low across all four subgroups, with a couple outliers in the male samples. Excluding the outliers in the non-COVID-19 male subgroup, they seem to be the most consistent, with the widest base compared to the other subgroups.

3.6 Summary Statistics Table

Table 1: Summary Statistics Stratified by ICU Status

ICU Status	n	Sex (male)	Disease status (COVID-19)	Age	A1BG Gene Expression	Charlson Score
No	60	33 (55%)	50 (83.3%)	58.7 (17.8)	0.44 (0.39)	3.15 (2.46)
Yes	66	41 (62.1%)	50 (75.8%)	63.5 (14)	0.29 (0.16)	3.82 (2.5)

Figure 6: Summary table of number of male samples, COVID-19 positive samples, average and standard deviation of age, A1BG gene expression, and Charlson score stratified by ICU status.

The summary table (Figure 6) is stratified by ICU status. The n indicates the sample size in each group, 60 samples were not in the ICU, while 66 were in the ICU. The number and percent of males in each group are also listed, as well as the number and percent of COVID-19 positive samples in each ICU subgroup. The average and standard deviation of the ages, A1BG expression values, and Charlson scores for each ICU subgroup are also calculated and displayed. There are slightly more males that were admitted to the ICU.

The prevalence of COVID-19 participants were lower in the ICU-admitted sample than the non-admitted samples. The average age and Charlson score was higher in the ICU-admitted participants, while the A1BG expression was lower compared to non-ICU-admitted participants.

References

[1] El Atab, O., Gupta, B., Han, Z., Stribny, J., Asojo, O. A., and Schneiter, R. Alpha-1-b glycoprotein (a1bg) inhibits sterol-binding and export by CRISP2. 107910.

[2] Kolde, R. *pheatmap: Pretty Heatmaps*, 2025. R package version 1.0.13.

[3] Overmyer, K. A., Shishkova, E., Miller, I. J., Balnis, J., Bernstein, M. N., Peters-Clarke, T. M., Meyer, J. G., Quan, Q., Muehlbauer, L. K., Trujillo, E. A., He, Y., Chopra, A., Chieng, H. C., Tiwari, A., Judson, M. A., Paulson, B., Brademan, D. R., Zhu, Y., Serrano, L. R., Linke, V., Drake, L. A., Adam, A. P., Schwartz, B. S., Singer, H. A., Swanson, S., Mosher, D. F., Stewart, R., Coon, J. J., and Jaitovich, A. Large-scale multi-omic analysis of COVID-19 severity. 23.

[4] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[5] Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. *dplyr: A Grammar of Data Manipulation*, 2023. R package version 1.1.4.

[6] Wickham, H., Hester, J., and Bryan, J. *readr: Read Rectangular Text Data*, 2024. R package version 2.1.5.

[7] Xie, Y. *Dynamic Documents with R and knitr*, 2nd ed. Chapman and Hall/CRC, Boca Raton, Florida, 2015. ISBN 978-1498716963.

[8] Zhu, H. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2024. R package version 1.4.0.